# A New Method for Side-Chain Conformation Prediction Using a Hopfield Network and Reproduced Rotamers

**HIDETOSHI KONO and JUNTA DOI***

*Bioinformation Engineering Laboratory, Department of Biotechnology, The University of Tokyo, 1-1-1, Yayoi, Bunkyo-ku, Tokyo 113, Japan*

*Received 29 August 1995; accepted 12 December 1995*

## ABSTRACT

We present a new side-chain prediction method based on energy minimization using a Hopfield network, focusing on the buried residues of proteins. In this method, the network is composed of automata assigned to each rotamer to restrict side-chain conformational space. We reproduced a rotamer library that enabled us to more widely cover the space for side-chain conformations than those previously produced. The accuracy of the side-chain modeling was estimated by three standards: root mean square deviations (rmsds) between the modeled and the crystal structures, the percentages of correctly predicted side-chain torsion angles, and the percentages of correctly predicted hydrogen bonds. The average rmsd for buried side chains of 21 proteins was 1.10 Å. The value was almost always improved relative to the previous works. The percentage of side-chain $\chi_1$ angles for buried residues was 87.3%. By considering the hydrogen bond energy, the average percentage of correctly predicted hydrogen bonds rose from 33% without hydrogen bond energy to 52% with the bond energy. We applied this method to homology modeling, where the protein backbone used to predict side-chain conformations deviates from the correct conformation, and could predict side-chain conformations as correctly as those using the correct backbones. 

## Introduction

We propose here a new method for predicting side-chain conformations using a Hopfield network.[1] In this article the prediction quality is focused on that of buried residues because the root mean square deviation (rmsd) for overall residues depends partly on that for exposed residues that are often flexible and cannot be definitely determined by X-ray crystallography while the buried residues are usually well determined with good electron density.

* Author to whom all correspondence should be addressed.



CCC 0192-8651 / 96 / 141667-17

Homology modeling is one of the most successful methods for predicting tertiary structures[2]; however, there is no generally accepted method for predicting side-chain conformations. Summers et al. reported that residues of similar size and electrostatic character tend to adopt the same conformation in an equivalent position between homologous proteins.[3] However, they indicated that the conformations of identical side chains in homologous structures were not always similar. Methods were then presented to search topologically similar segments in the data base and to select the best segment by energy minimizations.[4-7] However, if there are no highly homologous proteins, these methods have been found to break down.

On the other hand, the methods for side-chain prediction are presented that depend not on the homology but on the energy minimization. It is impossible to search all side-chain conformations to find the best combination because of their considerable freedom of torsion angles. The complexity then needs to be reduced to a manageable level by simplification. Ponder and Richards found that distributions of side-chain torsion angles had high peaks by analyzing highly resolved crystal structures and summarized them as a rotamer library.[8] Tuffery et al. then reproduced the rotamer library with an increasing number of 50 well-resolved crystallographic protein structures and with the tools of statistical cluster analysis.[9] Even if the rotamer libraries are used, it is impossible to exhaustively search the available space for side-chain conformation. Bruccoleri and Karplus[10], and Wilson et al.[11] repeated the energy minimization for one or a few residues to optimize side-chain conformations. Eisenmenger et al. predicted side-chain conformations based on side chain–backbone interactions.[12] Alternatively, in a nonexhaustive manner, methods such as a Monte Carlo algorithm,[13] a genetic algorithm,[9] and the mean field theory[14] were presented to simultaneously minimize the conformations of all side chains based on rotamer libraries. Desmet et al. applied a "dead end elimination" theorem to eliminate many possible rotamers,[15] but the effectiveness depends on protein size and crystal packing. Without a rotamer library approximation, Lee and Subbiah optimized side chains by randomly moving the side-chain torsion angles by 10°.[16] During a random change of a side-chain conformation in the above energy minimization methods, the change is difficult to accept because the conformation of the side chain can be influenced by those of neighboring residues in well-packed proteins. Using these minimization methods, in general, the final conformations are not always identical in each run; however, little attention has been given to the following point.

In general, the prediction accuracy for exposed residues was poorer and consequently for overall residues. Interactions with the solvent must be taken into account for exposed residues. Wilson et al. added an approximate solvation term to a molecular mechanics force field and obtained similar results for the exposed residues of four proteins.[11] All previous methods but one did not consider the solvent; the accuracy for the exposed residues seems to be superficial. The solvation effects are implicitly included in the rotamer libraries during the statistical derivation.

We expressed side-chain conformations using a rotamer library to restrict side-chain conformational space. We reproduced the library by applying a statistical cluster analysis to 100 proteins extracted from the Protein Data Bank (PDB),[17] reflecting the recent increase in protein structural data. In this study an energy minimization method using a Hopfield network was developed to predict optimal side-chain conformations. The potential energy we used included only van der Waals energies and hydrogen bond energies. Electrostatic energies were ignored because the energies could not be correctly calculated in the absence of a solvent. In our minimization method, an automaton with one interior state of a continuous value from zero to unity is assigned to each rotamer; then the automata were connected with each other to form a network. The automata update their interior state to decrease the network energy as the automata gradually become influenced by other automata. The network structure can properly express the process where residues determine their conformations by their surrounding residues. For crambin and trypsin inhibitor, we predicted side-chain conformations 100 times starting from various initial network states. The conformations of buried residues were then found to be more reliable to determine than those of exposed residues. A side-chain rmsd for buried residues was 1.10 Å, which was almost always better than that in the previous work, and the overall 21 protein side-chain rmsd was 1.73 Å. The percentage of side-chain $\chi_1$ angles for buried residues appeared to be better than any other work and the value was 87.3%, although the percentage for overall was at a similar level. In the predicted structures, 52% of the hydrogen bonds in the crystal structures could

be reproduced correctly by considering the hydrogen bond energy, the assessment of which was introduced for the first time. In homology modelings of a hen lysozyme from a human lysozyme and vice versa, the inherent deviations in the backbones slightly affected the overall rmsd.

## Materials and Methods

We predict side-chain conformations from a given backbone structure using an energy minimization based on a kind of Hopfield network. The network is composed of automata assigned to each rotamer. Every automaton is connected to another to form a network with a connective weight according to the mutual interaction energies. The network energy decreases monotonically as the automata interact with each other (Fig. 1). When the network energy converges, side-chain conformations are determined definitely.

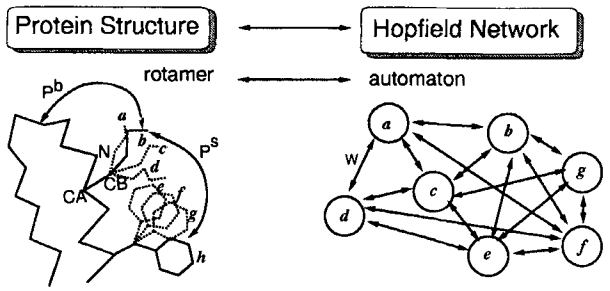


**FIGURE 1.** The architecture of the network. $P^b$ and $P^s$ are interactions between the backbone and rotamers and between rotamers, respectively. Every rotamer is assigned to an automaton (circle in the right figure). When the network converges, only one interior state of the automata of a residue site (for example, one of automata $a$–$d$ corresponding to rotamers $a$–$d$ in the left figure) becomes 1 (= output in the network) while the rest of the automata equal zero. $W$ denotes the connective weight between automata $a$ and $d$ calculated using eq. (3).

### ROTAMER LIBRARY

The prediction of side-chain conformations is reduced to a combinatorial problem when side-chain conformations are assigned to discrete rotamers. Tuffery et al. produced a revised rotamer library using 50 highly resolved proteins.[9] There are some side-chain conformations that cannot be covered with their library. Tertiary structural data increased thereafter due to the efforts of many crystallographers and NMR analysts. We then proposed a revised rotamer library using 100 current proteins with a statistical cluster analysis. Side-chain torsion angles are defined by the IUPAC-IUB

**TABLE I.**
**100 Proteins Used for Producing Rotamer Library.**

| Su[a] | PDB Code | Su | PDB Code | Su | PDB Code | Su | PDB Code | Su | PDB Code |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1YCC | 151 | 1PHH | 475 | 4PTP | 745 | 1HOE | 1183 | 2PAB |
| 10 | 2CDV | 156 | 2SOD | 486 | 1CSEE | 817 | 1XY1 | 1190 | 1UTG |
| 11 | 2CCY | 163 | 1FNR | 490 | 9PAP | 833 | 1GCN | 1257 | 2LTN |
| 13 | 256B | 186 | 8ATC | 494 | 1PSG | 836 | 1PPT | 1260 | 9WGA |
| 20 | 3B5C | 196 | 3CLA | 495 | 2TMN | 839 | 3INS | 1334 | 1ABP |
| 21 | 2CPP | 238 | 2AAT | 517 | 3BLM | 855 | 1TNF | 1335 | 2GBP |
| 23 | 1FD2 | 247 | 2YHX | 527 | 1PYP | 857 | 1NXB | 1381 | 1BRD |
| 24 | 1HIP | 249 | 1PFK | 578 | 2CTS | 860 | 1SN3 | 1446 | 1FC2 |
| 25 | 5RXN | 294 | 3PGK | 595 | 2CA2 | 861 | 2MLT | 1464 | 1ETU |
| 28 | 1TRX | 302 | 3ADK | 607 | 1WSY | 878 | 1CRN | 1561 | 3GAP |
| 30 | 1PAZ | 359 | 1RHD | 637 | 1YPI | 884 | 1ACX | 1590 | 2WRP |
| 33 | 4FXN | 364 | 1BP2 | 647 | 3PGM | 902 | 1I1B | 1605 | 2CRO |
| 34 | 8ADH | 413 | 2SNS | 661 | 2TS1 | 913 | 2RHE | 2113 | 2PLV |
| 39 | 6LDH | 419 | 2RNT | 683 | 2GLS | 920 | 2HLA | 2125 | 2RSP |
| 43 | 3ICD | 420 | 7RSA | 718 | 5PTI | 931 | 2MHR | 2135 | 3HVP |
| 65 | 1GD1 | 423 | 2TAA | 720 | 2OVO | 932 | 1UBQ | 2211 | 3HMG |
| 101 | 3GRS | 431 | 1LZ1 | 722 | 7API | 1012 | 1CTF | 2324 | 2TMV |
| 135 | 2CYP | 434 | 1L01 | 732 | 1CSE | 1090 | 1GCR | 2357 | 2STV |
| 136 | 7CAT | 471 | 5CPA | 738 | 4CPA | 1111 | 4CPV | 2358 | 4SBV |
| 140 | 1GP1 | 474 | 2SGA | 741 | 4SGB | 1180 | 1RBP | 2420 | 2GN5 |

[a] Superfamily index.

Commission recommendations.[18] Our rotamer library included almost all rotamers of previous libraries, adding some significant new rotamers. In our library, torsion angles for residues, Lys and Arg, were analyzed up to $\chi_4$ angles, whereas in their library, they were analyzed only up to $\chi_3$ angles. Our library included more rotamers for each residue than theirs, so that it could cover more conformational space for side chains.

## DATABASE USED

To avoid biasing, we classified the PDB (rel.58) data using the superfamily indices of IDEAS.[19] Some proteins were classified by cross references of Swiss-Prot (rel.22) when they were not found in the IDEAS. We selected the most highly resolved structure in every set of the superfamily. As a result, 100 proteins were selected from the PDB (Table I).

## CLUSTER ANALYSES

Our cluster analysis was applied to 17 residue types. Three residues, Pro, Gly, and Ala, and residues with missing or ambiguous atoms were typically excluded from consideration. The cluster analysis for an example of Asn is as follows.

1. Initial centers $a_i$ are allocated in a $\chi_1 - \chi_2$ plane, where $a_i$ denotes the $i$th center.
2. All data points of torsion angles are assigned to their nearest center.
3. Each $a_i$ is reassigned to the averaged coordinates of the points captured by the center $a_i$.
4. When the distance between a center $a_i$ and its nearest center $a_j$ is less than the distance $d = 20°$, and when the standard deviations of the two centers $a_i$ and $a_j$ do not exceed 20°, the two cluster sets are merged into one. Its center point is reassigned to the gravity center of the two centers $a_i$ and $a_j$. If at least one $\chi$ angle differs by more than 20° from that of a rotamer, its atom position deviates more than 0.5 Å and the side chain is considered as deviant from a geometric aspect.[20]
5. Procedures 2–4 are iterated until all standard deviations of cluster sets become less than 20° or until the iteration reaches a maximum time, which is set at 100 in this work.

Side chains of Cys, Ser, Thr, and Val were analyzed in a $\chi_1$ space, and those of Asn, Asp, His, Ile, Leu, Phe, Trp, Tyr, Lys, and Arg were in a $\chi_1 - \chi_2$ plane. For the residues Gln, Glu, and Met, the side chains were then analyzed in a $\chi_3$ space and for Lys and Arg in a $\chi_3 - \chi_4$. For Arg, the $\chi_5$ angles were fixed at 0°. For residues with only a $\chi_1$ angle, we gave 20 initial centers for every 18° interval and for those with $\chi_1$ and $\chi_2$ angles, $20 \times 20$ centers for every 18°. For the residues Phe, Tyr, and Asp, we considered their planarity of the $\chi_2$ angle and for Glu, the $\chi_3$ angle. The resulting rotamers are shown in Table II. For Ser, Thr, Val, Cys, Phe, and Tyr, the rotamers produced were less than 10 and for Asn, Glu, Gln, Met, Lys, and Arg, more than 20. The cluster analysis start with sufficient numbers of initial centers, and then the cluster centers close to each other are merged into one. In the process, each cluster center is not affected by data points that are far from it.

We assessed the rotamer libraries, using the 100 proteins, by the following criterion,

$$\sqrt{\sum_i (\Delta\chi_i)^2} \leq 20°, \qquad (1)$$

where $\Delta\chi_i$ denotes the $i$th torsion angle error between a correct rotamer and the best possible rotamer. In the space of torsion angles, the percentages of conformations covered with our library increased greatly from those of Tuffery et al.,[9] especially for Ser, Asn, Asp, Ile, and His (Table III). In their library, the percentages for residues with more than $\chi_3$ angles, such as Met, Glu, Gln, Lys, and Arg, were 15.8–45.3% for 10–16 rotamers. In our library, the values increased 25.6 to 54.5%, covering more space with 10 or less rotamers. To cover more than 60% of all conformations for Lys and Arg, we needed 96 and 69 rotamers. It is notable that the accuracy is insufficient for the two residues, although the rotamer representation is effective in the sampling of torsion angle space.

## ENERGY MINIMIZATION

We find an optimal combination of rotamers from a given backbone by minimizing interaction energies between the side chains and between the backbone and side chains. At first, an automaton is assigned to each rotamer. Then the automaton is connected with another to form a Hopfield network.[1] The network is characterized by a graded output, deterministic decision, and continuous activation in time. An automaton has a parameter

**TABLE II.**
**Reproduced Rotamer Library Analyzed by Statistical Cluster Analyses.**

| Arg 148[a] | | | | | | Lys 163 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Num[b] | $\chi_1$ | $\chi_2$ | $\chi_3$ | $\chi_4$ | $\sigma$[c] | Num | $\chi_1$ | $\chi_2$ | $\chi_3$ | $\chi_4$ | $\sigma$ |
| 41 | −65.1 | −177.9 | 179.1 | 175.1 | 37.4 | 107 | −72.7 | −179.8 | 179.5 | 176.8 | 29.1 |
| 32 | −68.5 | −168.0 | −69.1 | −90.7 | 31.3 | 79 | −177.6 | −171.9 | −177.1 | 178.5 | 31.0 |
| 32 | −66.4 | −176.9 | −178.0 | −97.1 | 32.0 | 57 | −51.3 | 177.0 | 174.9 | 176.6 | 31.3 |
| 28 | −65.6 | −178.9 | −64.6 | 170.2 | 28.7 | 43 | −177.6 | 155.0 | 179.3 | −177.1 | 36.5 |
| 26 | −67.6 | −179.8 | 171.7 | 97.7 | 31.3 | 35 | −62.0 | −60.2 | −175.9 | −176.3 | 38.7 |
| 23 | 179.0 | 176.7 | −177.6 | −179.8 | 32.7 | 34 | −177.1 | 75.7 | −179.9 | 173.1 | 36.0 |
| 21 | −178.0 | −179.5 | 67.3 | 83.5 | 26.0 | 32 | −60.4 | −146.5 | 178.5 | 168.5 | 33.1 |
| 19 | −54.6 | −75.8 | −58.2 | 175.1 | 34.5 | 31 | −92.8 | 144.4 | −178.3 | −172.0 | 40.0 |
| 18 | 179.0 | 179.4 | −179.5 | −87.5 | 35.5 | 29 | 57.0 | −175.4 | 177.9 | −179.1 | 38.2 |
| 17 | 177.5 | 175.2 | −62.5 | 101.2 | 38.9 | 27 | −58.0 | −91.2 | 179.0 | 173.3 | 38.6 |
| Glu 66 | | | | | | Gln 105 | | | | | |
| 161 | −68.9 | 175.8 | −14.7 | | 23.7 | 55 | −67.1 | 173.1 | −12.4 | | 22.8 |
| 82 | −72.9 | 172.8 | 39.6 | | 25.5 | 38 | −67.8 | 175.0 | 59.2 | | 29.0 |
| 70 | −66.4 | −62.8 | −56.4 | | 21.9 | 32 | −66.7 | −57.2 | −48.6 | | 26.3 |
| 68 | −72.0 | 178.4 | −75.9 | | 28.3 | 32 | −66.7 | 174.4 | −82.0 | | 23.8 |
| 65 | −176.4 | −175.3 | 10.6 | | 22.5 | 28 | −62.8 | 176.7 | −159.6 | | 24.0 |
| 48 | −68.4 | −72.1 | −5.4 | | 25.1 | 27 | −176.5 | 66.3 | 43.8 | | 20.3 |
| 43 | −174.1 | −167.7 | −47.2 | | 26.6 | 24 | −179.9 | 167.1 | −6.8 | | 22.3 |
| 33 | −60.9 | 78.6 | 15.2 | | 26.8 | 21 | −64.2 | 173.5 | 143.3 | | 23.2 |
| 32 | −173.8 | 69.3 | 46.5 | | 26.1 | 19 | −71.9 | −56.5 | 94.6 | | 39.6 |
| 30 | 164.9 | 155.6 | 17.3 | | 26.7 | 18 | 171.4 | −157.2 | −37.9 | | 27.4 |
| Met 73 | | | | | | Asn 22 | | | | | |
| 37 | −66.0 | 172.3 | 70.8 | | 22.6 | 130 | −66.6 | −56.7 | | | 18.2 |
| 29 | −64.8 | 173.2 | −178.5 | | 26.7 | 111 | −74.2 | −15.5 | | | 21.0 |
| 28 | −58.5 | −72.9 | −72.0 | | 23.3 | 80 | −63.3 | 124.4 | | | 18.7 |
| 25 | −69.2 | 178.8 | −78.8 | | 27.0 | 68 | −166.7 | 51.3 | | | 21.0 |
| 19 | −175.2 | −176.7 | 70.0 | | 20.0 | 67 | −79.1 | 167.5 | | | 23.8 |
| 19 | −74.9 | −54.7 | −70.2 | | 25.1 | 61 | −167.8 | 2.6 | | | 19.9 |
| 15 | −171.2 | 69.8 | 75.7 | | 23.3 | 52 | −165.1 | −50.4 | | | 24.1 |
| 14 | −171.6 | −171.4 | −177.9 | | 25.8 | 48 | −175.0 | −109.3 | | | 22.7 |
| 14 | −61.3 | −76.0 | 164.3 | | 23.9 | 48 | −83.5 | −104.9 | | | 29.1 |
| 11 | −174.6 | −174.7 | −71.7 | | 15.5 | 40 | −164.6 | −161.4 | | | 24.6 |
| Asp 13 | | | | | | Cys 4 | | | | | |
| 337 | −73.0 | −12.6 | | | 17.2 | 212 | −66.3 | | | | 14.7 |
| 212 | −62.7 | −53.7 | | | 21.7 | 107 | −177.7 | | | | 12.4 |
| 119 | −171.0 | −16.3 | | | 19.0 | 44 | 69.8 | | | | 10.1 |
| 113 | −178.7 | 33.3 | | | 21.8 | 18 | 37.6 | | | | 16.3 |
| 107 | 63.2 | 1.0 | | | 18.1 | | | | | | |
| 75 | −82.1 | 59.2 | | | 29.1 | | | | | | |
| 72 | −177.1 | 80.6 | | | 20.5 | | | | | | |
| 60 | 60.0 | −56.2 | | | 23.1 | | | | | | |
| 56 | −137.0 | 7.1 | | | 22.7 | | | | | | |
| 54 | −139.3 | −69.1 | | | 26.5 | | | | | | |
| His 20 | | | | | | Ile 23 | | | | | |
| 72 | −60.9 | −94.9 | | | 20.8 | 437 | −63.7 | 169.2 | | | 12.9 |
| 51 | −59.0 | −61.8 | | | 14.6 | 150 | −55.8 | −66.0 | | | 20.9 |
| 44 | −61.6 | 94.1 | | | 16.2 | 104 | 61.5 | 168.2 | | | 14.9 |
| 43 | 176.5 | 73.0 | | | 17.1 | 90 | −54.0 | −153.7 | | | 21.2 |
| 34 | −66.8 | 157.9 | | | 17.8 | 86 | −67.8 | 132.9 | | | 21.0 |
| 30 | 62.4 | 80.3 | | | 23.8 | 79 | −173.7 | 166.8 | | | 18.5 |
| 23 | −169.1 | −105.8 | | | 20.9 | 30 | −161.1 | 63.9 | | | 22.0 |
| 23 | 63.6 | −96.2 | | | 20.8 | 29 | −20.9 | 162.5 | | | 21.2 |
| 22 | −174.7 | −74.5 | | | 18.3 | 23 | 110.6 | −173.7 | | | 28.6 |
| 22 | −153.3 | 49.6 | | | 24.1 | 20 | −121.5 | −170.5 | | | 26.1 |

**(Continued)**

**TABLE II.**
**(Continued)**

| Arg 148[a] | | | | | | Lys 163 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Num[b] | $\chi_1$ | $\chi_2$ | $\chi_3$ | $\chi_4$ | $\sigma$[c] | Num | $\chi_1$ | $\chi_2$ | $\chi_3$ | $\chi_4$ | $\sigma$ |
| | | Leu 15 | | | | | | Phe 6 | | | |
| 749 | −66.0 | 175.1 | | | 19.9 | 352 | −66.5 | −82.4 | | | 22.5 |
| 377 | 177.5 | 66.0 | | | 18.7 | 269 | −179.6 | 78.8 | | | 21.0 |
| 122 | −102.5 | 28.8 | | | 20.4 | 83 | 61.4 | −89.4 | | | 21.8 |
| 105 | −138.0 | 59.1 | | | 20.5 | 75 | −72.1 | −8.0 | | | 30.3 |
| 83 | −139.6 | −147.5 | | | 27.5 | 34 | −162.5 | 12.4 | | | 33.6 |
| 75 | −77.3 | 83.3 | | | 28.9 | 5 | 77.7 | 14.4 | | | 32.7 |
| 71 | −158.4 | 147.3 | | | 30.1 | | | | | | |
| 38 | −84.0 | −92.1 | | | 26.1 | | | | | | |
| 37 | −68.5 | −27.8 | | | 26.5 | | | | | | |
| 28 | 50.9 | 70.1 | | | 31.0 | | | | | | |
| | | Ser 6 | | | | | | Thr 4 | | | |
| 394 | −62.1 | | | | 13.2 | 640 | −60.0 | | | | 17.5 |
| 380 | 59.8 | | | | 10.6 | 475 | 61.3 | | | | 15.0 |
| 316 | −176.8 | | | | 15.8 | 148 | −162.9 | | | | 21.1 |
| 129 | 97.3 | | | | 16.9 | 51 | 125.8 | | | | 21.1 |
| 113 | −109.2 | | | | 18.5 | | | | | | |
| 77 | 9.1 | | | | 19.2 | | | | | | |
| | | Trp 14 | | | | | | Tyr 7 | | | |
| 86 | −65.5 | 100.3 | | | 21.9 | 270 | −66.6 | −89.2 | | | 18.3 |
| 45 | −178.0 | 81.0 | | | 16.5 | 166 | 179.0 | 87.4 | | | 20.2 |
| 28 | 62.8 | −94.3 | | | 15.6 | 123 | −64.0 | −51.0 | | | 20.1 |
| 28 | −73.6 | −4.3 | | | 18.2 | 105 | −176.4 | 55.5 | | | 21.4 |
| 27 | −172.3 | −100.0 | | | 18.7 | 89 | 66.5 | −86.9 | | | 19.7 |
| 21 | −74.6 | −92.8 | | | 26.5 | 28 | −89.4 | 31.1 | | | 26.3 |
| 17 | 62.1 | 87.3 | | | 13.5 | 15 | 57.9 | 48.6 | | | 31.3 |
| 11 | 156.6 | −99.6 | | | 21.9 | | | | | | |
| 9 | −164.1 | 32.5 | | | 25.5 | | | | | | |
| 7 | −110.2 | 112.2 | | | 21.6 | | | | | | |
| | | Val 4 | | | | | | | | | |
| 967 | 175.2 | | | | 14.1 | | | | | | |
| 304 | −63.5 | | | | 16.0 | | | | | | |
| 163 | 76.4 | | | | 21.3 | | | | | | |
| 104 | 7.0 | | | | 21.1 | | | | | | |

[a] The final centers by the cluster analyses.
[b] The data points belonging to a center.
[c] Standard deviation of points belonging to a center.

that is mathematically called the interior state value. The interior state of an automaton is changed in response to the input value, and the automaton outputs a value according to the interior state. In the method, the output value is made equal to that of the interior state. The connective weight between automata is determined considering the interaction energies between automata and the constraint of the network. The convergent flow of the network to stable states depends on the connective weight and on the asynchronous interrogation of each automaton. The times of interrogation of each automaton are independent of the times at which other automata are interrogated. The interior states of the automata are updated one by one in random order on a discrete time scale. The cost of calculations is proportional to the square of the number of rotamers.

The interior state of the automaton has one value. The value continuously changes between 0 and 1 by receiving the mean interactions of connected automata. By repeating the procedure, only one automaton gradually changes its interior state to 1 at each residue site, while the rest of the automata reach zero. This means that every rotamer for a target residue site is gradually determined whether it is proper or not. In this way, conformations of surrounding residues affect the

**TABLE III.**
**Comparison of Rotamer Libraries.**

| Residue | Tuffery et al.[9] Cover (%)[a] | Num[b] | This work Cover (%) | Num |
|---|---|---|---|---|
| | | $\chi_1$ Angle | | |
| Cys | 86.4 | 3 | 89.2 | 4 |
| Ser | 68.2 | 3 | 83.4 | 6 |
| Val | 77.9 | 3 | 81.3 | 4 |
| Thr | 75.4 | 3 | 78.2 | 4 |
| | | $\chi_1 + \chi_2$ | | |
| Tyr | 74.6 | 6 | 87.2 | 7 |
| Trp | 72.0 | 7 | 82.6 | 10 |
| Ile | 66.4 | 5 | 80.0 | 10 |
| Asp | 46.3 | 3 | 77.9 | 10 |
| Phe | 70.5 | 3 | 77.8 | 6 |
| Leu | 64.1 | 6 | 76.8 | 10 |
| His | 54.3 | 6 | 70.8 | 10 |
| Asn | 35.0 | 4 | 60.1 | 10 |
| | | $\chi_1 + \chi_2 + \chi_3$ | | |
| Met | 45.3 | 10 | 60.8 (54.5)[c] | 20 (10) |
| Glu | 35.6 | 10 | 60.3 (54.6) | 30 (10) |
| Gln | 30.1 | 10 | 60.4 (43.1) | 29 (10) |
| | | $\chi_1 + \chi_2 + \chi_3 + \chi_4$ | | |
| Lys | 29.5 | 16 | 60.0 (26.5) | 96 (10) |
| Arg | 15.8 | 11 | 60.0 (25.6) | 69 (10) |

[a] A percentage of side-chain conformations covered with rotamers.
[b] Number of rotamer.
[c] Numbers in parentheses refer to cover rates with the best 10 rotamers (in Table II).

target conformation. One constraint is imposed to make only one rotamer at each residue site. When the interior state takes a binary value, that is 0 or 1, the minimization with the network is similar to such methods as a Monte Carlo algorithm, a simulated annealing algorithm, and a genetic algorithm.

We introduce the network energy in the following form,

$$E = \frac{A}{2}\left\{ \sum_{X=1}^{m} \sum_{Y \neq X}^{m} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} V_{Xi} V_{Yj} \left( P^s_{XiYj} + P^{sh}_{XiYj} \right) + \sum_{X=1}^{m} \sum_{i=1}^{n_X} V_{Xi} \left( P^b_{Xi} + P^{bh}_{Xi} \right) \right\} + \frac{B}{2}\left( \sum_{X=1}^{m} \sum_{i=1}^{n_X} V_{Xi} - 1 \right)^2 + \frac{B}{2} \sum_{X=1}^{m} \sum_{i=1}^{n_X} V_{Xi}(1 - V_{Xi}), \quad (2)$$

where $A$ and $B$ are positive empirical constants, $m$ is the number of residues, $n_X$ is the number of rotamers at a residue site $X$, and $X_i$ denotes a rotamer $i$ at a residue site $X$. In the above equation, the first term corresponds to the energies of van der Waals and hydrogen bonds. Energies $P^b$ and $P^s$ are the van der Waals energies between the backbone and a side chain and between the two side chains, and $P^{bh}$ and $P^{sh}$ are the hydrogen bond energies between the backbone and a side chain and between side chains. The second term is the topological constraint by which only one rotamer is determined at each residue site. The third is a term added due to a mathematical need for the substitution of $V_i^2$ with $V_i$.

When this energy $E$ is minimized using the network, the connective weight $W_{XiYj}$ between automata $Xi$ and $Yj$ and the excitation bias $\theta_{Xi}$ are determined as follows,

$$W_{XiYj} = -A\left(P^s_{XiYj} + P^{sh}_{XiYj}\right)(1 - \delta_{XY}) - B\delta_{XY}(1 - \delta_{ij}) \quad (3)$$

$$\theta_{Xi} = \tfrac{1}{2}B - \tfrac{1}{2}A(P^b_{Xi} + P^{bh}_{Xi}), \quad (4)$$

where $\delta_{ij}$ is 1 if $i = j$ and is 0 otherwise. These coefficients are obtained by comparing eq. (2) with the following equation,

$$E = \frac{1}{2} \sum_{X=1}^{m} \sum_{Y=1}^{m} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} W_{XiYi} V_{Xi} V_{Yj} - \sum_{X=1}^{m} \sum_{i=1}^{n_X} V_{Xi} \theta_{Xi}. \quad (5)$$

If the coefficients in eqs. (3) and (4) are determined like these, the network energy is guaranteed to decrease monotonically by the asynchronous update of each automaton.[21] The interior state of the automaton is given as follows,

$$V_{Xi}(t+1) = f(u_{Xi}(t)) = \frac{1}{2}\left(1 + \tanh\frac{u_{Xi}(t)}{u_0(\text{step})}\right), \quad (6)$$

where

$$\text{step} = \sum_{X}^{m} (mn_X)^2 \times t, \tag{7}$$

$$u_{Xi}(t) = \sum_{Y=1}^{m} \sum_{j=1}^{n_X} W_{XiYj} V_{Yj}(t) + \theta_{Xi}. \tag{8}$$

In the eq. (6) the parameter $u_0$ denotes an excitation parameter that regulates the extent of change in the interior state. The input of the automaton $u_{Xi}$ is the sum of the interior values of all connected automata. The network energy decreases monotonically as the interior value of the automaton is updated asynchronously. To avoid trapping relatively high local minima, we decrease the bias $u_0$ in accordance to the following expression,[22]

$$u_0(\text{step}) = u_0(0)\left(1 - \frac{\text{step}}{\tau}\right). \tag{9}$$

When $u_0$ is large enough in the first stage of a calculation, the interior states of all automata acquire a value of around 0.5. When the step goes forward and the $u_0$(step) is small enough, the interior states become nearly 0 or 1. In this way, a side-chain conformation is expressed as the average conformation of all rotamers, and then it gradually reaches a certain conformation.

In our calculation procedure, the initial value of the automaton interior state is given in the following form,

$$V = \frac{1}{n_X} \pm \Delta, \tag{10}$$

where $\Delta$ is a constant random number between 0.0 and 0.1. The parameters $A$ and $B$ were set to 1.0 and $10 \times A$, respectively. These parameters were empirically determined by balancing between the structural energies and the constraint. The initial bias $u_0(0)$ was set at 5.0, and the attenuation rate $\tau$ was 100. The calculation was cut off when the total step reached 100. The larger the attenuation coefficient $\tau$ is, the better the network functions. However, it is known that further improvement of the network decreases.[22] Therefore, we determined the above values $u_0(0)$ and $\tau$ by balancing between the calculation time and the network performance.

## ENERGY CALCULATION

Coordinates of side chains were constructed using the standard bond length and bond angles of the AMBER 3A parameters.[23] Our consideration was limited to interaction energies: van der Waals energies $P^s$ and hydrogen bond energies $P^{sh}$ between side chains, and those $P^b$ and $P^{bh}$ between the backbone and side chains. We judged that there was a contact between atoms; those energies greater than 100 kcal/mol were truncated to this maximal value. The electrostatic energy was omitted because it could not be estimated exactly. The energetic costs of altering bond lengths, bond angles, and dihedral angles could be neglected because rotamers with the idealized internal geometry of the atoms were used for the side chains. Cysteine residues linked by disulfide bonds were treated to be known in advance and their coordinates were fixed. The van der Waals parameters were those of the AMBER 3A.[23] Hydrogen bonds are a major component of tertiary structures.[24] The stereochemistry and energetics of this weak but essential chemical bond have been well studied. Its bond energy is estimated to be 2–10 kcal/mol.[25] We introduced the hydrogen bond energy based on the hydrogen bond geometrical criteria of Stickle et al.[26] They enable one to detect hydrogen bonds without generating hydrogen atoms and to take the rotation of hydrogen atoms into account. In these criteria, hydrogen bond radii were about 10% larger than the corresponding van der Waals' radii. Then the energy is as follows,

$$P^h = A'\left\{\left(\frac{r_0}{r}\right)^{12} - B'\left(\frac{r_0}{r}\right)^{10}\right\} \times \cos(\theta_1 - \alpha)\cos(\theta_2 - \beta)\cos(\theta_3)\cos(\theta_4), \tag{11}$$

where $r$ is the distance between atoms, $r_0$ is the sum of the two atom radii, D denotes a donor and A an acceptor. Angles $\theta_1$ and $\theta_2$ are the D-A-AA angle and the DD-D-A angle where DD and DD' denote donor antecedent atoms and AA and AA' acceptor antecedent atoms. Angles $\alpha$ and $\beta$ are optimal values for these angles, which are extracted from the literature.[26] Angle $\theta_3$ is an angle between the D-DD-DD' plane and the A-D-DD plane, $\theta_4$ between the A-AA-AA' plane and the D-A-AA plane. The minimum hydrogen bond energy depends on parameters A' and B', which were set at 20.0 and 1.2, respectively, based on the examinations of previous studies on hydrogen bonds.[26–30]

## ASSESSMENT OF CONVERGENCE

To assess the convergence of calculation, we define the entropy in the folded state at each

residue site as,

$$S = -R\sum_{i}^{n} p_i \ln(p_i), \qquad (12)$$

where $p_i$ is the probability of a rotamer $i$ at each residue site, $n$ is the total conformation number, and $R$ is the gas constant. The probability $p_i$ is calculated from the predicted rotamer distributions from 100 minimization runs starting with various initial states. The values $S$ vary from 0 to $R \ln(n)$. The large $S$ indicates that a side-chain conformation is hard to determine. The lowest bound of 0 in the $S$ indicates that a side-chain conformation always converges into the same one in the 100 runs.

### HOMOLOGY MODELING

Generally, the tertiary backbone trace is only known for a sequence homologue of the structure to be predicted. The coordinates of the backbone atoms (N, $C_\alpha$, C, O, and $C_\beta$) are to be fixed. The fix is one of the constraints in this article. In principle, it is possible to involve backbone fluctuations; however, that makes the combinatorial number vast beyond calculation. Homologous proteins have significant shifts in backbone positions that allow many more alternative packing arrangements, and the experiment has shown that mutations can also be accommodated in this way. No existing searching method has been able to take such flexibility into account, but this limitation must be acknowledged at present.

To test the accuracy of side-chain prediction in homology modeling, we chose two different but sequence-homologous structures of lysozyme, those from hen (6LYZ) and from human (1LZ1). The protein sequences were aligned using the program CLUSTAL_V.[31] The backbone of A was used to predict the side chains of B and vice versa. Insertions as well as gaps were omitted.

## Results and Discussion

### SIDE-CHAIN PREDICTION FOR CRAMBIN

We used crambin to examine two parameters in the network. This protein has 46 residues, including 12 buried residues. Network energy trajectories of crambin for three initial biases $u_0(0) = 10.0$, 5.0, and 0.1 with a constant attenuation coefficient $\tau = 100$, are shown in Figure 2a. The figure shows that the smaller the initial bias is, the faster the network energy converges. When the initial bias is small, the network state is trapped in local minima. On the other hand, when the initial bias is large, the state is hardly trapped in local minima, although it takes many steps to converge. We determined that if bias $u_0(0)$ was more than 2.0, the network energy was sufficiently minimized.

The trajectories for three attenuation rates, $\tau =$ 10.0, 50.0, and 100.0, with a constant bias $u_0 = 5.0$, are shown in Figure 2b. For a large $\tau$, the bias is reduced slowly according to the eq. (9). The more slowly the bias is reduced, the lower the final network energy is without trapping in local minima. On the other hand, the faster the bias is

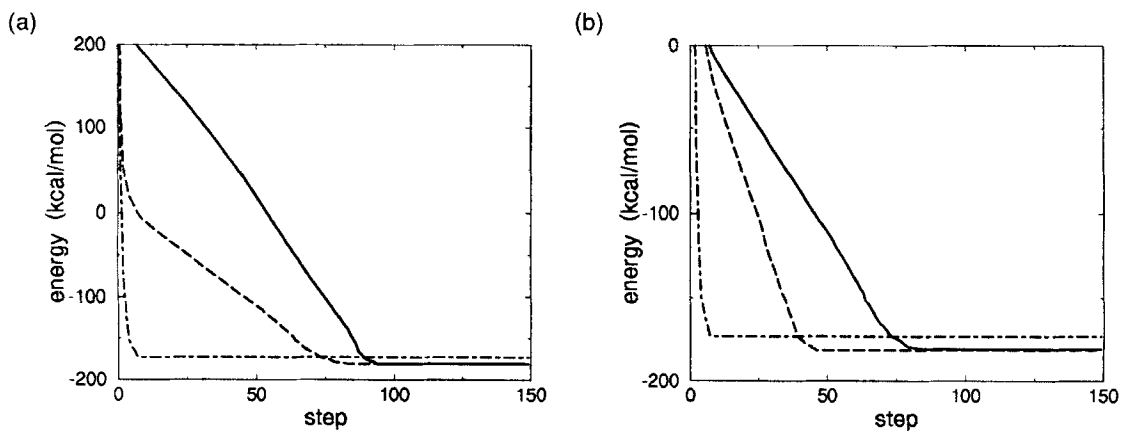


**FIGURE 2.** Network energy trajectories of crambin for two parameters, bias ($u_0$) and attenuation coefficient ($\tau$). (a) (—) $u_0 = 10.0$; (···) $u_0 = 5.0$; (-···) $u_0 = 0.1$. $\tau = 100.0$ for all. (b) (—) $\tau = 100.0$; (···) $\tau = 50.0$; (-···) $\tau = 10.0$. $u_0 = 5.0$ for all.

reduced, the faster the network converges. However, the network state is easily trapped in local minima. Considering the trade-off between the performance of the network and the calculation time, the initial bias $u_0(0)$ and the attenuation coefficient of the bias $\tau$ were set at 5.0 and 100.0, respectively. The maximum step was set at 100 considering the convergence flow of the network in the precalculations.

We predicted side-chain conformations 100 times, starting from various initial states, to determine the energy convergence. The energies of minimized structures deviated from −177.3 to −177.6 kcal/mol with a standard deviation of 0.04. This shows that each run converges to similar low energy structures. The entropies for buried residues, calculated using eq. (12), were smaller than those for exposed residues (Fig. 3).

This shows that buried residues almost always converge into the same conformation while conformations of exposed residues were hard to converge into the same conformation. Scattering of overall rmsds depends on that of exposed residues. For exposed residues, interactions with the solvent must be considered in the calculation. These residues exposed to the solvent fluctuate so that their conformations do not coincide with the conformation in the crystal structure.[32] For trypsin inhibitor, we also calculated 100 times and obtained results similar to those for crambin (data not shown).

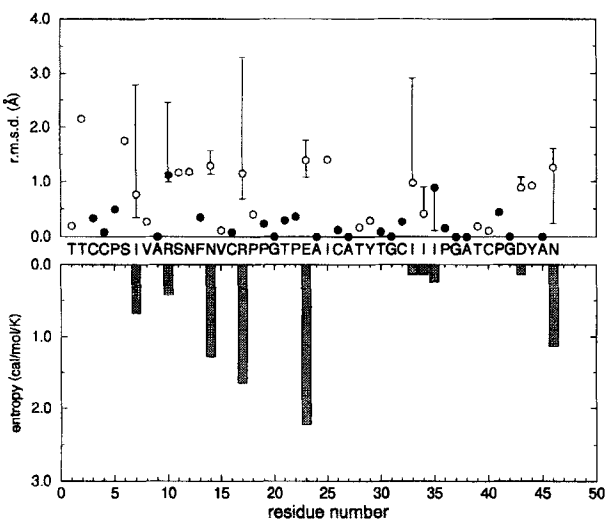


**FIGURE 3.** Mean, minimum, and maximum rmsds (top) and side-chain entropies (bottom) for each residue position from 100 runs for crambin using Rot80. Buried residues and fixed residues Gly, Pro, and Ala are shown by filled circles and otherwise by open circles.

The calculation time depends on the the size of a protein, and it is proportional to about the square of the size of a protein. For crambin with 46 residues, it took 42 CPU s with a workstation HP712 (122 SPECfp92) to predict side-chain conformations.

It is expected that the larger the number of rotamers involved in the prediction becomes, the better the quality of side-chain prediction is. To assess the size of the rotamers for each residue type, we then used four libraries, which were Rot10, Rot20, Rot40, and Rot80 (Table IV). The library Rot10 was composed of at most 10 rotamers for each residue type. For the residue type that had less than 10 rotamers, all rotamers tabulated were used. For example, 6 rotamers were used for Phe and 4 for Thr. In the same way, Rot20, Rot40, and Rot80 were composed of at most 20, 40, and 80 rotamers, respectively.

For crambin, we predicted side chains using the four libraries. The calculation was performed 100 times for each library. Results are shown in Table V. The average of overall side-chain rmsd for 100 times was best for Rot80 and its lowest rmsd was 0.76 Å. If side chains were modeled using rotamers with the best possible fit in the 100 runs, the overall rmsds could decrease to 1.22 Å for Rot10 and 0.70 Å for Rot80. If side chains were modeled using rotamers with the best possible fit rotamers in the libraries, the overall rmsds could decrease to

**TABLE IV.**
**Number Rotamers Used for Four Libraries.**

| | Libraries | | | |
|---|---|---|---|---|
| Residue | Rot10 | Rot20 | Rot40 | Rot80 |
| Cys | 4 | 4 | 4 | 4 |
| Val | 4 | 4 | 4 | 4 |
| Thr | 4 | 4 | 4 | 4 |
| Ser | 6 | 6 | 6 | 6 |
| Phe | 6 | 6 | 6 | 6 |
| Tyr | 7 | 7 | 7 | 7 |
| Asp | 10 | 13 | 13 | 13 |
| Trp | 10 | 14 | 14 | 14 |
| Leu | 10 | 15 | 15 | 15 |
| His | 10 | 20 | 20 | 20 |
| Asn | 10 | 20 | 22 | 22 |
| Ile | 10 | 20 | 23 | 23 |
| Glu | 10 | 20 | 40 | 66 |
| Met | 10 | 20 | 40 | 73 |
| Gln | 10 | 20 | 40 | 80 |
| Arg | 10 | 20 | 40 | 80 |
| Lys | 10 | 20 | 40 | 80 |

**TABLE V.**
**Results for Crambin from 100 Runs.**

| | Libraries | | | |
|---|---|---|---|---|
| | Rot10 | Rot20 | Rot40 | Rot80 |
| Mean rmsd (Å) | 1.31 | 1.33 | 1.24 | 0.95 |
| SD (Å) | 0.02 | 0.04 | 0.05 | 0.02 |
| Min. rmsd[a] (Å) | 1.25 | 1.23 | 1.18 | 0.76 |
| Max. rmsd[a] (Å) | 1.39 | 1.44 | 1.36 | 1.17 |
| Mean buried[b] (Å) | 0.75 | 0.83 | 0.71 | 0.62 |
| Min. buried (Å) | 0.75 | 0.65 | 0.65 | 0.48 |
| Max. buried (Å) | 0.92 | 0.83 | 0.71 | 1.14 |
| Mean exposed (Å) | 1.45 | 1.46 | 1.38 | 0.94 |
| Min. exposed (Å) | 1.38 | 1.34 | 1.30 | 0.83 |
| Max. exposed (Å) | 1.55 | 1.60 | 1.52 | 1.13 |
| Min. model[c] (Å) | 1.22 | 1.02 | 1.16 | 0.70 |
| Best model[d] (Å) | 0.70 | 0.47 | 0.47 | 0.33 |
| Entropy | | | | |
| All[e] | 0.22 | 0.27 | 0.31 | 0.25 |
| Exposed[e] | 0.32 | 0.38 | 0.43 | 0.33 |
| Buried[e] | 0.011 | 0.011 | 0.033 | 0.066 |

[a] The minimum and the maximum rmsds among 100 runs.
[b] The residues with less than 25% relative solvent accessible surface.
[c] The rmsd using rotamers with the best possible fit in the 100 solutions.
[d] The rmsd using rotamers with the best possible fit in the library.
[e] The mean entropy for all side chain except Gly, Pro, and Ala residues (cal / mol / K).

0.70 Å for Rot10 and 0.33 Å for Rot80. However, the obtained rmsds were larger than those values because there were unfavorable van der Waals contacts when side chains adopted crystallike conformations. It is difficult for the discrete rotamers, due to their nature, to avoid the unfavorable contacts by slightly rotating their torsion angles.

Our rotamer library (Rot10) covered larger side-chain conformational space than that previously reported (Table III), so that the results of side-chain prediction improved. Compared with the methods based on what is called energy minimization, we obtained the best rmsds for crambin, 0.66 Å for residues with less than 20% relative solvent accessible surface areas (Wilson et al.[11] 0.82 Å) and 0.75 Å for 25% (Tuffery et al.[9] 1.99 Å).

For crambin, if we use Rot80 instead of Rot10, 13 residues with more than 11 rotamers can increase the number of degrees of conformational freedom. As a result, five residues, Ser6, Leu7, Arg10, Arg17, and Asp43, improved in rmsd by 0.31, 0.24, 0.64, 3.52, and 2.02 Å (Fig. 4). In the residues, only Arg10 was buried and the rest were exposed to the solvent. This means that Rot10 almost always had a sufficient number of rotamers for buried residues. In general, the prediction will be reliable for proteins with more buried residues.

## PREDICTION ACCURACIES FOR 21 PROTEINS

We assessed our prediction method for 21 proteins with sizes ranging from 26 to 346. They were chosen for their highly resolved structures as well as for comparison with previously reported proteins. The backbone atoms (N, $C^{\alpha}$, C, O, and $C^{\beta}$) were fixed at first, and then the side-chain conformations were predicted. In the calculation, Rot10 was used to save memory.

The quality of side-chain prediction was evaluated both by side-chain atom rmsds and by percentages of torsion angles that were predicted correctly within $\pm 40°$ as described in previous work.[12] In addition to the above two criteria, we introduced hydrogen bond reproduction rates as a new criterion. These criteria were calculated over buried residues and also over all residues. The buried residues were defined as those having a solvent accessible surface area less than 25% of that of the same residue in the sequence Ala-X-Ala in an extended conformation. For ready comparison of rmsd with other studies, the rmsd calculation included $C^{\beta}$ atoms, though the $C^{\beta}$ positions were essentially fixed.

The accuracy of prediction was almost always better for just buried residues, with an average

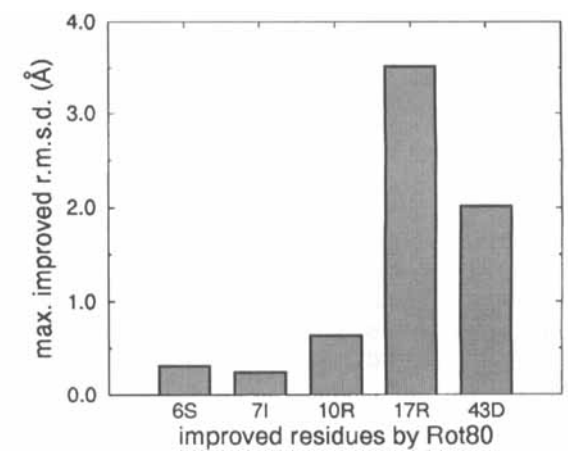


**FIGURE 4.** Improvement in rmsd using Rot80 instead of Rot10. These residues are all exposed ones except for Arg10.

rmsd improved to 1.10 Å from 1.73 Å of the average rmsd for overall side chains (Table VI). The accuracies of torsion angles were also better for buried residues, and 87% of $\chi_1$ and 74% of both $\chi_1$ and $\chi_2$, ($\chi_1 + \chi_2$), were correctly predicted. For overall residues, an average 74% of $\chi_1$ and 57% of $\chi_1 + \chi_2$ were correctly predicted. Especially for buried residues of three proteins, avian pancreatic polypeptide (1PPT), trypsin inhibitor (4PTI), and scorpion neurotoxin (1SN3), 100% of the $\chi_1$ angles were predicted correctly (Table VII).

Accuracies of side-chain prediction for different residue types are shown in Table VIII. In a comparison of side-chain torsion angles between predicted and crystal structures, we considered the rotational symmetry axis in Phe, Tyr, Asp, Glu, and Arg residues. The table shows that prediction accuracies of hydrophobic residues such as Val, Leu, Ile, Phe, Tyr, and Met were better than those of the hydrophilic type. For all residues except Glu, Cys, and Arg, prediction accuracies for buried residues became better. Serine residues were poorly predicted, because Ser is quite small and its position is usually determined by its ability to form hydrogen bonds. It should be noted that Ser residues with small side chains have a small rmsd in spite of the poor prediction of torsion angles. On the other hand, residues with large side chains such as Trp, His, and Tyr have large rmsds in spite of correct prediction of torsion angles.

Residues with bulky side chains such as Val, Thr, Phe, Tyr, Ile, His, and Trp, were predicted better. It was expected that residues with long side chains, such as Glu, Gln, Met, Lys, and Arg, would be poorly predicted; however, the expectation was not true of Met. In the proteins tested, 86% of the methionine residues appeared in the buried regions, thus they had many contacts with the surrounding atoms to restrict their conformations. As we can see in the Table III, a percentage of side-chain conformations covered with rotamers for Met is higher than that for Glu, Gln, Lys, and Arg. This means that Met residues take canonical conformation almost always, but Glu, Gln, Lys, and Arg do not. As a result, the average rmsd was small (1.09

**TABLE VI.**
**Accuracies of Side-Chain Prediction from Correct Backbone Coordinates.**

| PDB Code | Res (Å) | Size | Side Chain with H-B[a] (Å) | | Side Chain without H-B[b] (Å) | |
|---|---|---|---|---|---|---|
| | | | Buried | All | Buried | All |
| 2MLT | 2.0 | 26 | — | 1.22 | — | 1.24 |
| 1PPT | 1.37 | 36 | 0.69 | 1.83 | 0.69 | 1.57 |
| 1CRN | 1.5 | 46 | 0.75 | 1.30 | 0.91 | 1.67 |
| 2OVO | 1.5 | 56 | 0.99 | 1.51 | 1.00 | 1.63 |
| 4PTI | 1.5 | 58 | 0.85 | 1.80 | 0.97 | 2.07 |
| 1SN3 | 1.8 | 65 | 0.78 | 1.77 | 0.83 | 1.60 |
| 1CTF | 1.7 | 68 | 1.11 | 2.06 | 1.11 | 2.08 |
| 1UBQ | 1.8 | 76 | 0.80 | 1.96 | 0.86 | 2.04 |
| 1PCY | 1.6 | 99 | 0.99 | 1.68 | 1.5? | 1.92 |
| 5CYT | 1.5 | 103 | 1.37 | 2.05 | 38 | 2.12 |
| 1HMQ | 2.0 | 113 | 1.18 | 1.75 | 1.42 | 1.95 |
| 3RN3 | 1.45 | 124 | 1.02 | 1.81 | 0.99 | 1.88 |
| 7RSA | 1.26 | 124 | 0.89 | 1.68 | 1.53 | 2.13 |
| 6LYZ | 2.0 | 129 | 1.14 | 1.70 | 1.32 | 2.17 |
| 1LZ1 | 1.5 | 130 | 1.38 | 2.00 | 1.68 | 2.28 |
| 3FXN | 1.9 | 138 | 1.81 | 1.84 | 1.86 | 1.85 |
| 3TLN | 1.6 | 316 | 1.30 | 1.74 | 1.48 | 1.96 |
| 3APP | 1.8 | 323 | 1.10 | 1.28 | 1.27 | 1.43 |
| 2APR | 1.8 | 325 | 1.08 | 1.55 | 1.31 | 1.74 |
| 2LIV | 2.4 | 344 | 1.30 | 1.87 | 1.48 | 1.89 |
| 2LBP | 2.4 | 346 | 1.41 | 1.90 | 1.52 | 2.06 |
| Average | 1.7 | 153 | 1.10 | 1.73 | 1.25 | 1.87 |

[a] Backbone–side chain and side chain–side chain interactions with hydrogen bond energies.
[b] Backbone–side chain and side chain–side chain interactions without hydrogen bond energies.

**TABLE VII.**
**Percentage of Correctly Predicted $\chi$ Angles.**

| PDB Code | Size | $\chi_1$ (%) | | $\chi_1 + \chi_2$[a] (%) | | All $\chi$[b] (%) | |
|---|---|---|---|---|---|---|---|
| | | Buried[c] | All | Buried | All | Buried | All |
| 2MLT | 26 | — | 95.0 | — | 75.0 | — | 60.0 |
| 1PPT | 36 | 100.0 | 75.9 | 75.0 | 65.5 | 75.0 | 58.6 |
| 1CRN | 46 | 90.0 | 84.4 | 90.8 | 68.8 | 90.0 | 65.6 |
| 2OVO | 56 | 92.9 | 71.1 | 85.7 | 57.8 | 85.7 | 55.6 |
| 4PTI | 58 | 100.0 | 88.1 | 78.6 | 61.9 | 71.4 | 50.0 |
| 1SN3 | 65 | 100.0 | 73.5 | 89.5 | 53.1 | 84.2 | 49.0 |
| 1CTF | 68 | 85.7 | 67.4 | 78.6 | 50.0 | 78.6 | 34.8 |
| 1UBQ | 76 | 96.2 | 72.3 | 76.9 | 52.3 | 73.1 | 44.6 |
| 1PCY | 99 | 87.9 | 62.3 | 72.7 | 46.8 | 72.7 | 42.9 |
| 5CYT | 103 | 83.3 | 66.2 | 77.8 | 53.8 | 77.8 | 50.0 |
| 1HMQ | 113 | 90.9 | 70.1 | 68.2 | 50.5 | 63.6 | 46.4 |
| 3RN3 | 124 | 80.9 | 70.5 | 72.3 | 55.2 | 68.1 | 50.5 |
| 7RSA | 124 | 87.2 | 75.2 | 83.0 | 63.8 | 78.7 | 55.2 |
| 6LYZ | 129 | 89.1 | 76.7 | 69.6 | 56.3 | 65.2 | 48.5 |
| 1LZ1 | 130 | 92.3 | 81.6 | 78.8 | 63.1 | 75.0 | 55.3 |
| 3FXN | 138 | 68.6 | 70.4 | 52.9 | 47.8 | 52.9 | 43.5 |
| 3TLN | 316 | 84.0 | 73.4 | 69.5 | 57.8 | 67.9 | 54.1 |
| 3APP | 323 | 84.4 | 76.5 | 71.6 | 64.0 | 68.1 | 58.7 |
| 2APR | 325 | 84.8 | 77.0 | 73.1 | 63.4 | 71.7 | 60.5 |
| 2LIV | 344 | 75.7 | 62.5 | 64.0 | 47.4 | 59.6 | 39.8 |
| 2LBP | 346 | 72.6 | 64.7 | 57.0 | 44.2 | 51.9 | 38.9 |
| Average | 153 | 87.3 | 74.0 | 74.3 | 57.1 | 71.6 | 50.6 |

[a] $\chi_1 + \chi_2$ is defined as those side chains with both $\chi_1$ and $\chi_2$ correct. Side chains for which only $\chi_1$ exists are included in $\chi_1 + \chi_2$.
[b] All $\chi$ is defined as those side chains with all $\chi$ angles correct.
[c] Side chains with less than 25% of their extended surface accessible to the solvent.

Å) and the percentage of correctly predicted $\chi_1$ angles was high (84%).

Because the criteria of the prediction accuracy are different for each study, comparison of the results obtained here with those reported by others is not always fair. However, with this in mind, it is interesting to compare the results here with those of other workers (Table IX). Our method yielded the best overall rmsds for some proteins, but not for others. In general, our predictions for buried residues were almost always better than those of other workers. In comparison with the results of Tuffery et al.,[9] our rmsd for each residue type except Lys and Gln improved. In particular, the values for Phe, Trp, and His improved by more than 1.0 Å. Compared with Dunbrack and Karplus,[6] our rmsd values for Leu and aromatic residues improved by more than 0.6 Å. For Tyr, Phe, Val, and Ile, the percentages were as high as those of Eisenmenger et al.[12] In this way, our method achieved good predictions for hydrophobic residues because our network can properly express the environment where hydrophobic residues almost always appear in buried regions so that they can form extensive contacts with surrounding atoms of the backbone and other side chains. In our network, the target conformation is affected by mean interactions of the surrounding residues and gradually determined whether it is proper or not by continuous interior value of the assigned automaton. The use of continuous value works favorably in overcoming the multiple local minima in energy space.

## HYDROGEN BOND

A hydrogen bond is one of the major determinant factors in protein folding[24]. However, there was no previous discussion of the reproduction of the hydrogen bond. We studied the effect of the hydrogen bond energy by checking the percentage of the hydrogen bonds correctly reproduced in the predicted structure to the number of all hydrogen bonds in the crystal (Table X). The hydrogen bonds between the backbone and side chains and those between side chains were calculated. These bonds

**TABLE VIII.**
**Percentage of Correctly Predicted $\chi$ Angles for Different Residue Types.**

| Residue | Number | | rmsd (Å) | | $\chi_1$ (%) | | $\chi_1 + \chi_2$ (%)[a] | | All $\chi$ (%)[b] | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Buried[c] | All | Buried | All | Buried | All | Buried | All | Buried | All |
| Val | 150 | 200 | 0.36 | 0.43 | 89.3 | 86.0 | | | | |
| Leu | 147 | 189 | 0.88 | 0.91 | 85.7 | 83.6 | 62.6 | 59.3 | | |
| Ile | 134 | 168 | 0.67 | 0.72 | 86.6 | 85.7 | 62.7 | 60.1 | | |
| Ser | 67 | 225 | 0.82 | 0.95 | 52.2 | 46.2 | | | | |
| Thr | 89 | 197 | 0.69 | 0.67 | 70.5 | 70.6 | | | | |
| Asp | 68 | 201 | 1.15 | 1.37 | 73.5 | 62.2 | 39.7 | 27.9 | | |
| Asn | 44 | 168 | 1.47 | 1.56 | 68.2 | 61.3 | 34.1 | 34.1 | | |
| Lys | 17 | 192 | 1.37 | 2.12 | 70.6 | 53.6 | 47.1 | 31.8 | 29.4 | 13.5 |
| Glu | 32 | 132 | 1.67 | 1.80 | 59.4 | 63.6 | 40.6 | 34.8 | 25.0 | 19.7 |
| Gln | 39 | 135 | 1.40 | 1.91 | 76.9 | 68.9 | 74.4 | 48.9 | 33.3 | 16.3 |
| Arg | 12 | 93 | 1.89 | 2.56 | 66.7 | 61.3 | 58.3 | 46.2 | 18.3 | 16.7 |
| His | 30 | 43 | 0.95 | 1.20 | 93.3 | 88.4 | 60.0 | 48.8 | | |
| Phe | 95 | 111 | 0.64 | 0.69 | 97.9 | 96.4 | 87.4 | 84.7 | | |
| Cys | 66 | 76 | 0.32 | 0.40 | 93.9 | 89.5 | | | | |
| Trp | 32 | 39 | 1.26 | 1.50 | 90.6 | 87.2 | 75.0 | 69.2 | | |
| Tyr | 87 | 138 | 0.76 | 0.98 | 96.6 | 92.0 | 88.5 | 84.2 | | |
| Met | 37 | 43 | 0.97 | 1.09 | 89.2 | 83.7 | 81.1 | 74.4 | 64.9 | 58.1 |

[a] $\chi_1 + \chi_2$ is defined as those side chains with both $\chi_1$ and $\chi_2$ correct. Side chains for which only $\chi_1$ exists are included in $\chi_1 + \chi_2$.
[b] All $\chi$ is defined as those side chains with all $\chi$ angles correct.
[c] Side chains with less than 25% of their extended surface accessible to the solvent.

were detected by the program we implemented based on the definition of Stickle et al.[26]

The hydrogen bonds in buried regions denote those formed by either a donor or an acceptor with relative solvent accessibility of less than 25%. Percentages of correct reproduction without the hydrogen bond energy were 31% for buried regions and 33% for overall structures. With the energy, the percentages improved by 15% for the buried ones and 19% overall. The percentage for buried regions was poor because it was presumed that the number of candidates for hydrogen bonds were few and these candidates were difficult to predict correctly. The total number of hydrogen bonds in the predicted structures with the hydrogen bond energy was nearly equal to that of the crystal structures; however, the number without the hydrogen bond energy was 55%. The predicted structures with the hydrogen bond energy had some incorrect hydrogen bonds so that there were some cases where the rmsd was poorer than that without it, such as 1PPT and 3RN3 in Table VI.

**TABLE IX.**
**Comparison of Side-Chain rmsd in This Work with Those in Previous Works.**

| Method | Proteins | | Overall (Å) | | Method Common (Å) | | Network Common (Å) | |
|---|---|---|---|---|---|---|---|---|
| | Total | Common | Buried (Å) | All | Buried | All | Buried | All |
| Lee & Subbiah[16] | 9 | 3 | 1.3 | 1.8 | 1.5 | 2.0 | 1.1 | 1.6 |
| Tuffery et al.[9] | 15 | 9 | 1.5 | 1.8 | 1.6 | 1.8 | 1.0 | 1.8 |
| Holm & Sander[13] | 33 | 9 | 1.4 | 1.8 | 1.2 | 1.7 | 1.1 | 1.8 |
| Eisenmenger et al.[12] | 6 | 5 | 1.1 | 1.7 | 1.1 | 1.7 | 0.8 | 1.7 |
| Laughton[5] | 8 | 5 | 1.0 | 1.7 | 0.9 | 1.7 | 1.0 | 1.7 |
| Kohel & Delarue[14] | 30 | 18 | 1.4 | 1.9 | 1.3 | 1.9 | 1.1 | 1.7 |
| This work | 21 | | 1.1 | 1.7 | | | | |

**TABLE X.**
**Accuracies of Hydrogen Bond Prediction from Correct Backbone Coordinates.**

| PDB Code | Crystal Num[b] | With H-B[a] (%) | | | Without H-B (%) | | |
|---|---|---|---|---|---|---|---|
| | | Num | Buried[c] | All | Num | Buried | All |
| 2MLT | 3 | 6 | — | 100 | 4 | — | 67 |
| 1PPT | 2 | 11 | — | 100 | 4 | — | 50 |
| 1CRN | 12 | 11 | 44 | 36 | 9 | 44 | 42 |
| 2OVO | 26 | 28 | 50 | 62 | 15 | 43 | 46 |
| 4PTI | 19 | 18 | 50 | 61 | 10 | 40 | 47 |
| 1SN3 | 13 | 13 | 20 | 23 | 8 | 40 | 36 |
| 1CTF | 12 | 5 | 0 | 25 | 1 | 0 | 0 |
| 1UBQ | 24 | 18 | 40 | 38 | 15 | 33 | 32 |
| 1PCY | 31 | 31 | 59 | 55 | 18 | 28 | 28 |
| 5CYT | 27 | 37 | 73 | 63 | 24 | 59 | 46 |
| 1HMQ | 28 | 29 | 54 | 45 | 16 | 29 | 31 |
| 3RN3 | 52 | 59 | 61 | 60 | 40 | 29 | 28 |
| 7RSA | 51 | 57 | 59 | 57 | 35 | 26 | 26 |
| 6LYZ | 54 | 45 | 32 | 35 | 21 | 9 | 13 |
| 1LZ1 | 57 | 52 | 59 | 51 | 33 | 28 | 28 |
| 3FXN | 26 | 29 | 27 | 35 | 14 | 33 | 33 |
| 3TLN | 124 | 113 | 51 | 48 | 71 | 32 | 30 |
| 3APP | 125 | 118 | 48 | 47 | 70 | 28 | 31 |
| 2APR | 99 | 108 | 55 | 59 | 51 | 30 | 31 |
| 2LIV | 75 | 97 | 50 | 52 | 56 | 33 | 23 |
| 2LBP | 75 | 97 | 51 | 39 | 50 | 25 | 18 |
| Average | 45 | 45 | 46 | 52 | 27 | 31 | 33 |

[a] H-B denotes the results with the hydrogen bond energies.
[b] The number of hydrogen bonds between the backbone and side chains and between side chains.
[c] Hydrogen bonds involved either a donor or an acceptor with less than 25% of their extended surface accessible to solvent.

With the hydrogen bond energy, an average rmsd was 0.15 Å better than that without it; but this difference was slight. The wrong hydrogen pairs were observed in the predicted structures with hydrogen bond energy. The residues exposed to solvent often form hydrogen bonds with water molecules, but our calculations contained no water. This resulted in the hydrogen pairs being unwillingly searched in the protein molecule and then the wrong hydrogen bonds being formed.

In summary, hydrogen bond energy improves the reproduction of hydrogen bonds, although it is not directly reflected in the rmsd values.

## HOMOLOGY MODELING

The accuracy of side-chain prediction in homology modeling was tested using this method. We chose two different but homologous structures of the proteins, human lysozyme (PDB code 1LZ1) and hen lysozyme (PDB code 6LYZ). The two sequences were aligned using the CLUSTAL_V,[31] and the results are shown in Figure 5. The identity between them was 58%, and the rmsd between backbones was 0.89 Å. Residue 45 of 1LZ1 with no structural equivalence in 6LYZ was excluded from the analysis. The backbone of 1LZ1 was used to predicted the side-chain conformations of 6LYZ and vice versa.

Figure 6 shows the predicted and observed conformations using the backbone of 1LZ1 to predict the 6LYZ side-chain conformations. It was observed that residues exposed to the solvent some-

```
1LZ1  KVFERCELAR  TLKRLGMDGY  RGISLANWMC  LAKWESGYNT  RATNYNAGDR
6LYZ  KVFGRCELAA  AMKRHGLDNY  RGYSLGNWVC  AAKFESNFNT  QATN-RNTDG
      *** *****     ** * * *  ** ** ** *  ** **  **   ***    *
1LZ1  STDYGIFQIN  SRYWCNDGKT  PGAVNACHLS  CSALLQDNIA  DAVACAKRVV
6LYZ  STDYGILQIN  SRWWCNDGRT  PGSRNLCNIP  CSALLSSDIT  ASVNCAKKIV
      ****** ***  ** ***** *  **  * *     *****   *     * *** *
1LZ1  RDPQGIRAWV  AWRNRCQNRD  VRQYVQGCGV
6LYZ  SDGNGMNAWV  AWRNRCKGTD  VQAWIRGCRL
       * *  ***   ******   * *     **
```

**FIGURE 5.** Alignment of 1LZ1 and 6LYZ using the CLUSTAL_V program.[31] (*) identical residue.

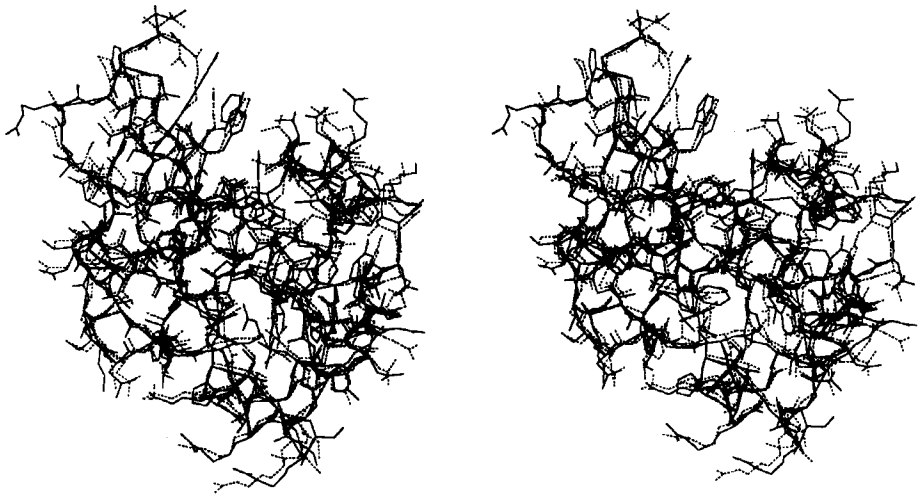

**FIGURE 6.** A stereo view of the (—) predicted and (···) crystal structures for hen lysozyme using the backbone trace of human lysozyme.

times deviated greatly, but, in general, conformations were predicted correctly. In 6LYZ modeling using the backbone of 1LZ1, the overall rmsd was 1.40 Å, and in the opposite modeling, the rmsd was 1.64 Å (Table XI). In this case, the overall rmsd involved backbone atoms. The former value was 0.50 Å better than that of Wilson et al.[11] (in the latter case, data were not shown in Wilson et al.). The percentages of torsion angles correctly predicted were better than those of previous studies.[11, 14]

## PREDICTION LIMITATIONS

Prediction for exposed residues was also poor in previously reported methods. This was true of the present work. Several environmental factors are likely to be responsible for the inaccurate prediction of surface side-chain conformation. The hydrophilic residues were charged residues so that they interacted electrostatically with other molecules such as solvent, substrates, and atomic packing in the crystal from neighboring molecules. In general, it is expected that side chains are stable if they take torsion angles of about $-60°$, $+60°$, or $180°$. There are $3^4 = 81$ stable conformations for Lys residues having four $\chi$ angles. The results of cluster analysis showed that 163 conformations (see Table II) were needed for Lys data instead of the 81 conformations. This suggests that Lys residues often assume noncanonical conformations. This is also true of Glu and Gln (Table II). Therefore, it is necessary to take intermolecular interactions into account. However, in buried regions, the present method is sufficiently valid because the

**TABLE XI.**
**Quality of Homology Modeling.**

| Target | Template | Size | Identity | Backbone rmsd (Å) | All atom rmsd (Å) | | $\chi_1$ (%) | | $\chi_1 + \chi_2$ (%) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Buried[a] | All | Buried | All | Buried | All |
| 6LYZ | 1LZ1 | 129 | 58 | 0.89 | 1.02 | 1.40 | 82.6 | 74.3 | 63.0 | 52.5 |
| 6LYZ | 6LYZ | 129 | 100 | 0.00 | 0.84 | 1.22 | 89.1 | 76.7 | 69.6 | 56.3 |
| 1LZ1 | 6LYZ | 129 | 58 | 0.89 | 1.38 | 1.64 | 82.7 | 75.2 | 63.5 | 60.2 |
| 1LZ1 | 1LZ1 | 130 | 100 | 0.00 | 1.02 | 1.44 | 92.3 | 81.6 | 78.8 | 64.1 |

[a] Side chains with less than 25% of their extended surface accessible to solvent.

side-chain conformations are considered to be determined mainly by intramolecular interactions.

## References


1. J. J. Hopfield and D. W. Tank, *Biol. Cybernet.*, **52**, 141 (1985).
2. T. L. Blundell, B. L. Sibanda, M. J. E. Sternberg, and J. M. Thornton, *Nature (Lond.)*, **326**, 347 (1987).
3. N. L. Summers, W. D. Carlson, and M. Karplus, *J. Mol. Biol.*, **196**, 175 (1987).
4. M. Levitt, *J. Mol. Biol.*, **226**, 507 (1992).
5. A. C. Laughton, *J. Mol. Biol.*, **235**, 1088 (1994).
6. R. L. Dunbrack, Jr. and M. Karplus, *J. Mol. Biol.*, **230**, 543 (1993).
7. C. A. Schiffer, J. W. Caldwell, P. A. Kollman, and R. M. Stroud, *Proteins: Struct. Funct. Genet.*, **8**, 30 (1990).
8. J. W. Ponder and F. M. Richards, *J. Mol. Biol.*, **193**, 775 (1987).
9. P. Tuffery, C. Etchebest, S. Hasout, and R. Lavery, *J. Biomol. Struct. Dyn.*, **8**, 1267 (1991).
10. R. E. Bruccoleri and M. K. Karplus, *Biopolymers*, **26**, 137 (1987).
11. C. Wilson, L. M. Gregoret, and A. Agard, *J. Mol. Biol.*, **229**, 996 (1993).
12. F. Eisenmenger, P. Argos, and R. Abagyan, *J. Mol. Biol.*, **231**, 849 (1993).
13. L. Holm and C. Sander, *Proteins: Struct. Funct. Genet.*, **14**, 213 (1992).
14. P. Koehl and M. Delarue, *J. Mol. Biol.*, **239**, 249 (1994).
15. J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters, *Nature (Lond.)*, **356**, 539 (1992).
16. C. Lee and S. Subbiah, *J. Mol. Biol.*, **217**, 373 (1991).
17. F. C. Bernstein, T. F. Koetzle, E. F. Williams, G. J. B. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
18. O. O. Hoffmann, W. E. Cohen, A. E. Braunstein, P. Karlson, B. Keio, W. Klyne, C. Liebeca, E. C. Slater, E. C. Webb, and W. J. Whelan, *Biochemistry*, **9**, 3471 (1970).
19. M. Kanehisa, *IDEAS User Manual*, Advanced Scientific Computing Laboratory, National Cancer Institute, 1988.
20. H. Schrauber, F. Eisenhaber, and P. Argos, *J. Mol. Biol.*, **230**, 592 (1993).
21. H. Kono and J. Doi, *Proteins: Struct. Funct. Genet.*, **19**, 244 (1994).
22. Y. Akiyama, *Trans. Inst. Electron. Inf. Commun. Eng.*, **NC90-40**, 73 (1990).
23. J. S. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, S. Alagona, Jr., G. Profeta, and P. Weiner, *J. Am. Chem. Soc.*, **106**, 765 (1984).
24. K. A. Dill, *Biochemistry*, **29**, 7133 (1990).
25. T. E. Creighton, *PROTEINS*, W. H. Freeman and Company, New York, 1993.
26. D. F. Stickle, L. G. Presta, K. A. Dill, and G. D. Rose, *J. Mol. Biol.*, **226**, 1143 (1992).
27. I. K. McDonald and J. M. Thornton, *J. Mol. Biol.*, **238**, 777 (1994).
28. A. Vedani and J. D. Dunitz, *J. Am. Chem. Soc.*, **107**, 7653 (1985).
29. E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.*, **44**, 97 (1984).
30. R. Taylor, O. Kennard, and W. Versichel, *J. Am. Chem. Soc.*, **105**, 5761 (1983).
31. D. G. Higgins, A. J. Bleasby, and R. Fuchs, *CABIOS*, **8(2)**, 189 (1992).
32. M. Billeter, A. D. Kline, W. Braun, R. Huber, and K. Wüthrich, *J. Mol. Biol.*, **206**, 677 (1989).